修士論文

学校非公式サイトにおける
有害語の極性判定に関する研究

平成　22 年度修了

三重大学大学院工学研究科
博士前期課程情報工学専攻

松葉　達明

要旨

「ネット上のいじめ」が新しい「いじめ」の形態として問題となっている.「ネット上のいじめ」とは，携帯電話やパソコンを通じてインターネット上のいわゆる学校非公式サイトの掲示板などにおいて，特定の子どもの悪口や誹謗・中傷を書込んだり，メールを送信するなどして，有害情報によるいじめを行うものである．これらの有害情報は，ネットパトロールにより監視されている．ネットパトロールとは，文字通り，学校非公式サイトの掲示板などを人手でつぶさにチェックを行う監視作業である．しかしながら，現状では，ネットパトロールにおける書込みの確認作業が最も負担が大きく，増大し続ける学校非公式サイトを監視するのは困難となる．そこで，本研究では学校非公式サイトの掲示板に書込まれる有害情報を検出するシステム構築を目指す.

まず，有害情報と無害情報，それぞれの書込み中の各単語を主な分析対象として，言語表現の分析をした．その結果，名詞,動詞,形容詞で,有害と無害の上位を占める単語には,品詞により出現傾向の違いが見られた．名詞では個人名や「バカ」などの誹謗中傷語や卑猥語が目立ち，動詞では「死ね」などの暴力誘発語，形容詞では「キモイ」などの誹謗中傷語が支配的であった,さらに，有害情報中の単語間の係り受け関係を調べたところ,特定の要素が組み合わされるという条件によって有害化する傾向が見られた．例えば,「性格が悪い」や「胸がでかい」などの有害表現は,「性格-悪い」,「胸-でかい」という係り受けで構成されている．しかし，その構成要素である「性格」や「胸」,「悪い」,「でかい」のみが単独で出現したとしても有害性を持たず，これらの要素が係り受け関係を持って共起することによって，はじめて有害性を持つのである．このような係り受け関係を持つ要素の組を有害性判定の素性として用いることは,有害表現の判別に大きく寄与すると考えられる.

次に，有害情報と無害情報の分類実験を行った.提案手法は,(１)有害情報候補単語列の抽出，(２)有害な極性を持つ単語の参照，(３)拡張 PMI-IR による有害表現判別，という三つのステップで処理を実施する.(1)では，有害情報かどうかを判定する要素を書込みから抽出する.この要素には，言語表現の分析結果から「名詞-名詞」,「名詞-動詞」,「名詞-形容詞」のいずれかの係り受け関係を持つ単語の組とした．(2)では，有害な極性を持つ単語を参照する．その有害な極性を持つ単語には,「きもい」,「死ね」などの 9 個の単語を参照した．(3)では，(1)で抽出した判定要素と(2)の有害な極性を持つ単語との関連度を算出する．つまり，この関連度が高ければ高いほど，書込みは有害情報の可能性が高いということを示している.

提案手法を用いて分類実験を行った．有害情報と無害情報を入力し，全ての書込みの関連度を算出して関連度順にランキングした.上位の有害情報候補単語列には,「不細工－顔」や「眉毛－濃い」などで占められており，誹謗中傷などの表現による有害情報が上手く取れていた．ランキングの上位 400 件以上を有害情報と判定した場合，適合率 0.83，再現率

0.32 という高い分類精度だった.下位の有害情報候補単語列には，人名や，住所，電話番号などの個人情報の流布による有害情報と意味不明な単語列で占められていた.これらの有害情報は,有害な表現と無害な表現の軸上に無い,中性的な表現である．よって，分類実験の結果から提案手法では,「悪い-女」や「うざい-先生」などの有害な表現と無害な表現の軸上にある有害情報は抽出できるが，人名や住所などの軸上に沿わない有害情報は抽出が難しいことがわかった．さらに，市販されている有害情報フィルタリングソフトを想定し，有害単語マッチング手法による比較実験を行った．その結果，有害単語マッチング手法でしか取れない有害情報はあるが，拡張 PMI-IR 手法の phrase の抽出規則を拡張すれば対応できるものであった．また，両方の手法で取れない個人情報の流布などは，個人情報の抽出に関する先行研究がなされており，その研究を応用すれば取得できると考えられる．

目次

第１章　　序論

1.1　研究の背景と目的

「ネット上のいじめ」が新しい「いじめ」の形態として問題となっている．「ネット上のいじめ」とは，携帯電話やパソコンなどを通じてインターネット上のいわゆる学校非公式サイトの掲示板などに特定の子どもの悪口や誹謗・中傷を書込んだり，メールを送信するなどして，有害情報によるいじめを行うもの[1]である．このような「ネット上のいじめ」では，短期間で深刻化するケースも多い上に，当事者は容易に被害者にも加害者にも成り得る．そのため，エスカレートを見過ごすと事件にまで発展する危険性があり，早期発見・早期対応に向けた取組みが急務である．

文部科学省は，「ネット上のいじめ」を手段や内容に着目して表 1.1 のように類型化している．このうち，学校非公式サイトの掲示板における「ネット上のいじめ」に注目してみる．学校非公式サイトの掲示板は，複数のユーザが相互に発言を行い，情報交換をする場である．このようなサイトでは，議論の食い違いや学校での諍いなどが発端となり，他のユーザが不快と感じる発言や特定個人を誹謗中傷する発言が書込まれるケースも頻繁に発生する[2].

これらの有害情報は，図 1. 1 のようにネットパトロールに基づいて対応されている．ネットパトロールとは，文字通り学校非公式サイトなどを人手でつぶさに書込み内容のチェックを行うことである．ネットパトロールによって有害であると判断された書込みについては，該当掲示板の管理人あるいはプロバイダに削除依頼がなされる．

表 1.1 文部科学省におけるネット上のいじめの類型化

1.　掲示板・ブログ・プロフでの「ネット上のいじめ」 (a). 掲示板・ブログ・プロフへの誹謗・中傷の書込み (b). 掲示板・ブログ・プロフへの個人情報無断掲載 (c). 特定個人になりすましたインターネット活動
2.　メールでの「ネット上のいじめ」 (a). メールによる特定の子どもに対する誹謗・中傷 (b).「チェーンメール」での悪口や誹謗・中傷 (c).「なりすましメール」での誹謗・中傷
3.　その他

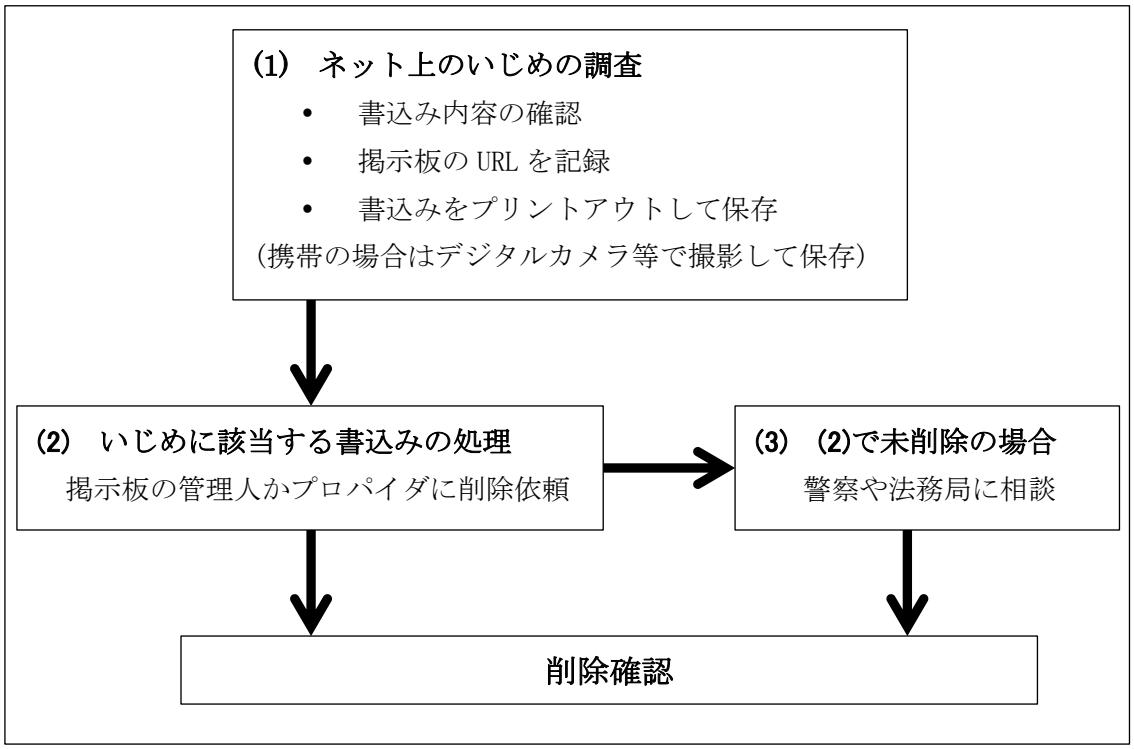

図 1.1　ネットパトロールの流れ

しかしながら，これらの活動は，教育委員会や学校の教職員や外部委託の情報教育アドバイザーがボランティアベースで行っている場合がほとんどである．掲示板などの書込みについても，書込み内容の印刷や携帯電話のカメラで画面を撮影するなどの処理を経た後に詳細な内容チェックを行うなど，気の遠くなるような作業を行っているのが現状である．また,ネットパトロール支援ツールとして,単語レベルで一致したものを検出する技術もあるが,単語レベルでの検出技術では,有害情報には該当しない多くの情報をも検出するなど,検出精度に問題がある.このような状況では，増大し続ける学校非公式サイト全てを監視し続けることは次第に困難となるであろうし，活動に取り組む人の健康や生活への影響も大きなものとなろう．

そこで本研究では，学校非公式サイトの掲示板に書き込まれる有害情報を検出するシステム構築を目指す．これにより，ネットパトロール活動の一部を自動化し，担当者の負担を軽減することができ，有害情報の早期発見・早期対応の支援にもつながる．

1.2　論文の構成

本論文は，全 8 章で構成されている．第 1 章の序論に続き，第 2 章では，ネットパトロールを支援するための基本的な考え方について述べる．

第 3 章では，本研究に関連する先行研究について述べる．

第 4 章では，有害情報の定義について説明する．また，作成した定義の評価実験について述べる．

第 5 章では，有害情報を抽出する前の準備について述べる．

第 6 章では，有害情報を抽出する提案手法について述べる．本手法は，拡張 PMI-IR による書込みと有害な単語との関連度によって抽出を行う．

第 7 章では，提案手法を用いた有害情報と無害情報の分類実験を行い，その結果について考察する．併せて，比較実験を行った．

第 8 章では，本研究に関する結論を述べる．

第２章　　基本的な考え方

本研究では,ネット上の掲示板やブログ等の自由に書き込めるスペースにおいて，特定個人の情報を流出したり誹謗中傷等を行い,被害者の実生活に悪影響を及ぼす情報を有害情報とする.有害情報の詳細な定義については，第4章で述べる．この有害情報の対処方法であるネットパトロールでは，掲示板の書込みを一つ一つ確認して,書込み内容の保存,削除依頼，削除の確認を行っている．現状では，無数に存在する書込みの確認が最も負担が大きく問題となっている.また,量の問題だけでなく，書込みによる被害者が存在するかどうか,第三者が判断するのは困難であり,疑わしい書込みは全てチェックする必要がある.以上を踏まえ,本研究ではネットパトロールにおける有害情報検出作業を支援する手法の構築を目指している.

現在,提案しようとする手法の概要(図2.1)について説明する．本手法は，(1)有害情報検出，(2)検出した有害情報の有害度によるランク付け，(3)有害情報可視化からなる．以下，各処理について述べる.

(1)有害情報検出

この処理では，上述したように疑わしい書込みは全て取り出すということを念頭に，有害情報を抽出する．また，その後の有害度によるランク付けを行うため，有害情報と無害情報の2値分類判定ではなく，有害度を考慮した閾値による分類手法が望ましい.

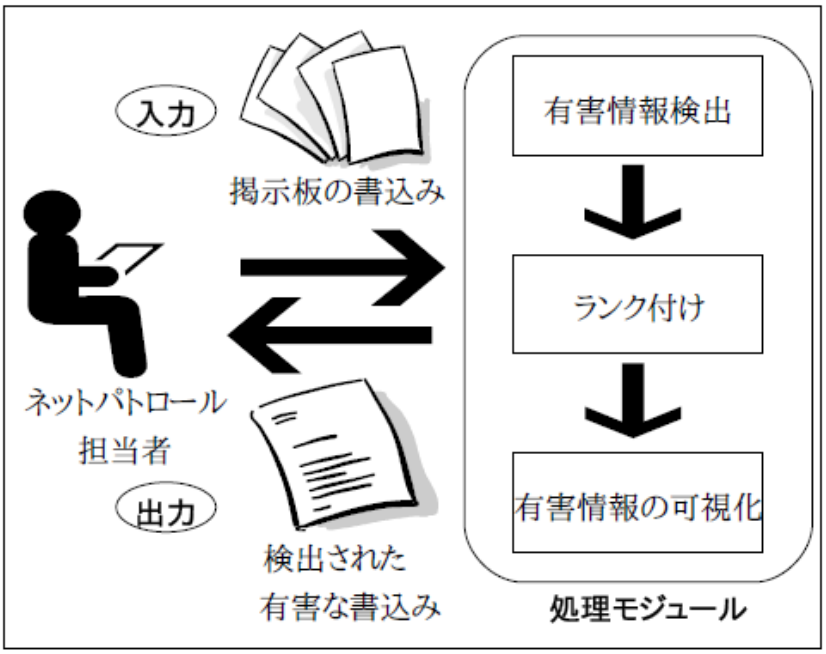

図 2.1　提案手法の概要

(2)有害度によるランク付け

抽出した有害情報に対し，早期に対処すべき有害度の高い有害情報を上位にランク付けする処理を行う．この有害度が高い有害情報とは，「明日，松葉を殴り飛ばす」のような，インターネット上だけで済まず，最悪の場合，事件になりうる有害情報を指す．つまり，より具体的にいじめ対象者を断定でき，より直接被害が及ぶ手段が記述されてあり，より近い日時でいじめを実行する記述を含む書込みが，有害度が高いと言える．

(3)有害情報の可視化

ネットパトロール担当者に必要な情報を提供し，使いやすいインターフェースを提供する．ランク付けした結果を提示し，ネットパトロール担当者の最終的な判断後，書込み内容，該当URLの保存などを自動化して，できる限りの労力を削減するのが望ましい．

本研究では，最も基本となる有害情報検出について研究を行った．

第３章　　関連研究

有害情報の抽出に関する先行研究としては，石坂ら[3]の研究と池田ら[4]の研究があげられる．

石坂らの研究では，巨大電子掲示板「2 ちゃんねる」[1]を対象とし，悪口表現辞書を構築している．石坂らは，悪口表現を「バカ」や「マスゴミのクズ」などの特定の他者に対して直接侮辱や誹謗中傷している単語，句と定義している．これらの悪口表現の使われ方，すなわち悪口表現に連接する単語の繋がりやすさを考慮し，周辺単語列から悪口表現を抽出することを試みている．句も，悪口表現の対象にしているので「バカな女」，「バカ美味い」のように，文脈によって有害と無害に分かれる表現にも対処できる．そのため，悪口表現辞書が構築できれば，単純に悪口表現が含まれているかどうかで有害情報を抽出することができる．しかし，悪口表現にのみ連接しやすい単語列の数は少なく，また悪口表現は定型的に存在するわけではないことを報告している．つまり，周辺単語列から悪口表現を抽出するのは困難であり，単語や句そのものが悪口かどうかを判断する必要がある．

池田らの研究では，人手で有害と無害に分けられた学習用文書を用いて，単語の出現頻度の偏りによる有害判定キーワードリストを構築している．文脈によって有害と無害に分かれる単語の問題については，単語の係り受け関係を利用して対処し，分類性能を向上させた．しかし，この手法では人手で学習用文書を作成する膨大な手間が問題となる．また，ウェブ文書では「爆破」と「爆ー破」のように少しだけ文字を変えた表現も多く，日々増え続ける新しい表現に，人手で学習用文書を作成していては次第に対応しきれなくなる．

本研究では，周辺単語列を考慮せず有害情報候補単語列を有害無害に判定する．また，ウェブ検索ヒット数を判定基準に利用するため，人手で学習用文書を構築する必要も無い．

[1] http://www.2ch.net/

第 4 章　　有害情報

4.1　　有害情報の定義

有害情報検出フェーズで，検出すべき有害情報の定義について説明する.有害情報の定義は，実用性を重視するため実際にネットパトロールに携わっている担当者の助言を受けて作成した.

まず，ネットパトロールにおいて対象の書込みは以下の3つのうちどれかに判断される.

- harmful：有害（削除）
- doubtful：有害（審議）
- normal：無害

「harmful:有害(削除)」は「明日，三重太郎を殺す」のように，具体的ないじめ対象者が挙げられ，対象者に直接被害が及びかねない早急に対処すべき書込みを指す.「doubtful:有害(審議)」は「大学でうざい奴」のような，付近にharmfulな書込みが現れる,またはそれ自体が削除対象となる可能性を含む，チェックしておきたい書込みを指す.normalは無害な書込みである.この3つに当てはめる有害情報の定義を表4.1に示す.

4.2　　定義の評価実験

ネットいじめに関与する書込みは個々の解釈の仕方によって大きく異なり，客観的な判断が非常に難しいという特徴がある．そのため，ネットパトロール担当者の助言を受けて作成した定義とはいえ，個々の解釈の仕方によって分類判定が分かれる可能性がある．そこで，曖昧な定義になっていないかを評価するために分類判定の実験を行った.

4.2.1　　実験手法

実際にネットパトロールによって収集された掲示板への書込みデータと，筆者らが独自に収集した書込みデータ(三重県域に限定されたサイトから収集したもの)2998件を対象に実験を行った．まず，有害・無害情報を含む書込みデータから,無作為に500件抽出する.そして，その書込みを検者に表4.1の定義を見ながら三つの分類に判別してもらう.判別してもらったデータは，CohenのKappa値(1)により分類結果にどの程度の差異があるのか見て,評価した.CohenのKappa値とは，単純に何パーセント一致し,何パーセント一致しなかったかという観測された一致率に対し,期待される一致率を考慮に入れたものである.

表 4.1 有害情報の定義

１．個人名の扱い

- 個人名そのものが記述されている　→　*harmful*

（例）「三重　太郎」，「太郎」

（メモ）「太郎って凄い人」のように肯定的な表現であっても対象とする.

- イニシャル、ニックネームが記述されている（個人名に準ずるものと判断）

（例）「三〇〇郎」,「なっちゃん」

  - 記述された当事者が特定できる記述　→　*harmful*
  - 記述された当事者が特定できない記述　→　*doubtful*

- 個人の所属や所有物などが記述されている

（例）「北見工大のテキスト情報処理研究室の准教授」

  - 記述された当事者が特定できる記述　→　*harmful*
  - 記述された当事者が特定できない記述　→　*doubtful*

２．個人情報の扱い

- 住所や連絡先などが記述されている

（例）「三重県津市一身田一二三」,「090-1234-5678」

  - 一般個人に関する個人情報の記述　→　*harmful*
  - 公共性の高い情報や公開情報の記述　→　*doubtful*

- 個人名や個人情報の書込みを誘導する記述がされている　→　*harmful*

（例）「うちの高校のヤリ〇ンって誰？」

（メモ）「ーのイケメンは誰？」のような肯定的な表現も対象とする.

- 個人に関する情報を提供する記述がされている

（例）「あいつ三重大で〇〇講義担当してる」,「そいつ確か三重大出身者」

  - 記述された当事者が特定できる記述　→　*harmful*
  - 記述された当事者が特定できない記述　→　*doubtful*

３．有害表現の扱い

- 誹謗中傷,暴力を誘発する言葉,猥褻な言葉が記述されている　→　*doubtful*

（例）「うざい」、「死ね」等の書込みを含む

（メモ）対象者が特定できない場合も含む

- 個人間の相互書込みの応酬（匿名も含む）　→　*doubtful*

（メモ）削除対象に発展する可能性がある

４．以上の条件に合致しない場合　→　*normal*

$$Cohen\text{の}Kappa\text{値} = \frac{P_o - P_c}{1 - P_c} \quad (1)$$

$$P_o = \text{観測された一致率}$$

$$P_c = \text{期待される一致率}$$

CohenのKappa値は0ー１を推移し，一般的に，値が0.41ー0.60の間ならば中等度の一致，0.61ー0.80の間ならば強い一致を示し，0.80を超える値をとる場合はほぼ一致していると考えられている．

4.2.2　実験結果

実験は成人男性5人，女性1人の計6人に行ってもらった．その結果，CohenのKappa値は2/3となり，検者同士の分類判定結果はかなりの一致を示した．

4.2.3　実験考察

実験の結果，検者同士の分類判定結果は強い一致度を示した．この結果から，個人の判断に依存する曖昧性はある程度排除されていると考えられる．しかし，「ニート」，「派遣」，「暴走族」などの誹謗中傷として扱うか否かでdoubtfulかnormalか，「○○もり，まえ○，○か○○」などの記述された当事者が特定できるか否かでharmfulかdoubtfulに判定するかで差異が現れていた．また，「三重県ってゴンズイ多いよね」という書込みがdoubtfulに判定されていることがあった．ゴンズイとは，魚の一種であり，本来ならnormalに判定される書込みである．これは，判定者には「ゴンズイは魚」という予備知識が無く，何か悪口の一種だと思いdoubtfulと判定した事例である．

第５章　　準備

まず，有害情報と無害情報，それぞれの書込みを構成する形態素を主な分析対象として言語表現の分析をした．形態素とは，それ以上分解したら意味をなさなくなるところまで分割して抽出された最小の文字列を指す．掲示板の書込みに対し形態素解析，係り受け解析を行うため，形態素解析に Chasen(ver2.3.3,日本語辞書 ipadic-2.7.0)[5]，係り受け解析に Cabocha(ver0.53)[6]を利用した.形態素解析,係り受け解析の解析例を表 5.1 に示す.これらの解析器は,既存の新聞記事など向けに作成された解析器であり，方言や若者言葉,表記のゆれなどの様々な表現方法が含まれる掲示板の書込みに対しては，解析精度が低くなってしまう．そこで，一定の解析精度を保つために有害単語の辞書登録と書込みデータの修正を人手で行った．以下，施した処理について詳述する.

表 5.1 各解析器の解析例

- **形態素解析**

  (入力):三重大学の松葉です

  (出力):三重大学　ミエダイガク　三重大学　名詞－固有名詞－組織
  　　の　　ノ　　の　　助詞－連体詞
  　　松葉　　マツバ　　松葉　　名詞－一般
  　　です　　デス　　です　　助動詞　特殊・デス　基本形

- **係り受け解析**

  (入力):三重大学の松葉です

  (出力)<ORGANIZATION>三重大学</ ORGANIZATION >の－D
  　　松葉です

5.1　　有害単語の辞書登録

有害情報を構成する一つの要素として「キモい」や「チャラ男」などの特有の単語(有害単語)が含まれる.有害単語は,既存の解析器に用意された日本語辞書には含まれておらず,未知語判定，もしくは解析失敗をしてしまう可能性が高い．しかし，有害単語は直接誹謗中傷する単語も多く，書込みに含まれるだけで有害情報となりうる重要な要素である．これらの単語を解析失敗してしまうと，有害情報と無害情報の分類は著しく困難となる．そこで，有害単語を人手で収集して整理し，あらかじめ形態素辞書に登録した．以下に作業の手順を示す.

1. 学校非公式サイトの掲示板から書込みデータを収集する.
2. 収集した書込みデータを読み，有害情報を含む書込みを抽出する.
3. 抽出した書込みから有害単語を抽出し，リストに加える.

上記手順に従って，有害単語の辞書登録を行った．まず，掲示板の書込みデータは，4.2.1 節の 2998 件の書込みを対象した．そこから，表 4.1 の有害情報の定義に当てはまる書込みを抽出したところ，1508 件の有害情報を含む書込みが得られた．有害単語であるかどうかの判断は，「ネット上のいじめ」マニュアル[1]に記載された「サイト・スレッドの書込み類型化」(表 5.2)に基づいて行った.

表 5.2 サイト・スレッドの書込み類型化

・ 「キモイ」「うざい」などの誹謗・中傷の語が含まれる
・ 「性器の俗称」など猥褻な語が含まれる
・ 「死ね」「消えろ」「殺す」など暴力を誘発する語が含まれる

以上の作業の結果，239 件の有害単語を辞書登録した．登録した有害単語は，解析器の辞書登録定義に従って，品詞情報，見出し語,優先コスト,読み,発音,活用形の情報を付与している．優先コストは，値が低ければ低いほど，その単語から優先して形態素解析を行う．有害単語は，重要な有害情報構成要素なので取りこぼしは避けたい．よって，最優先で解析されるように，他の全ての単語より低い 300 という値から，文字列のバイト数を引くことによって，最長一致法による最優先解析をするようにした．登録例を表 5.3 に示す.

表 5.3 辞書登録例

(品詞(形容詞 自立))((見出し語(キモい 294))(読み キモイ)(発音 キモイ)(活用型 不変化型))
(品詞(名詞 一般))((見出し語(ぶちゃいく 290))(読み ブチャイク)(発音 ブチャイク))

5.2 掲示板書込みの標準化

電子掲示板は，誰でも手軽に書込むことができ，閲覧する対象者も不特定多数のため，正しい文章，標準語で書かないくだけた書込みが多い．その他にも，書込み途中で誤送信し途中で文が切れた書込みや，意味不明の言葉を書き殴った書込みなど，電子掲示板には様々な書込みが存在する．それらのくだけた書込み例を表 5.4 に示す.

表 5. 4 くだけた書込み例

・ どっかに教えてもらったらしいいいいいー
・ 今も見とるんやろな．でてこやんかな．
・ サンクス．ちょっといってみるわ．

このような書込みは，形態素解析器で解析ミスをしてしまう可能性が高い．そこで，表 5. 5 にある標準化項目に従い，用意した掲示板の書込みデータ 2998 件中 1430 件を標準化した．

しかし，書込みの中には「あ」,「ん dsjc」,「ニャニャ～～yooooooooo!!」などの意味不明なもの,「タイヤが滑ってツツ～ドン！」,「お肉をジュージュー」などの擬態語を含むものなど，修正できない書込みが含まれていた．これらの書込みを除外することも考えられるが，実際に存在する書込みであり，同じ書込みに有害情報が混在する場合もあるため，書込みの除外は避けた．

表 5. 5 標準化項目一覧

1. 「若者言葉」 (例)つか → というより
2. 「掲示板特有の表現」 (例)ksk → 加速
3. 「略語」 (例)ほむぺ → ホームページ
4. 「方言」 (例)気いつけや → 気を付けろ
5. 「他言語表現」 (例)そーりー → すみません
6. 「隠語」 (例)円 → 援助交際
7. 「当て字」 (例)臭 L1 → 臭い
8. 「冗長的な表現」 (例)教えてあげましょうかぃ？ → 教えてあげましょうか？
9. 「日本語ミス」 (例)殺害しよーぜ → 殺害しようぜ
10. 「変換ミス」 (例)主，シネ → 主，死ね
11. 「形態素解析ミス」 (例)(笑) → (笑い) ※「笑」が人名と解析されてしまう

第 6 章　　提案手法

提案手法には，PMI-IR を応用した．PMI-IR とは，ウェブ検索ヒット件数を利用した共起度判定手法である[7]．まず，PMI は式(2)のように 2 つの *word* の関連度を示す．*p(word)* は，ドキュメント中の *word* の出現頻度を指し，*p(word₁ & word₂)*はドキュメント中の *word₁* と *word₂*の共起頻度を指す．

$$PMI(word_1, word_2) = log_2\left\{\frac{p(word_1 \quad \& \quad word_2)}{p(word_1)p(word_2)}\right\} \qquad (2)$$

そして，PMI-IR はIR(ウェブ検索ヒット件数)によって式(3)のように与えられる．*hits(word)* は *word* を検索単語としたときのウェブ検索ヒット件数を指す．*hits(word₁ & word₂)*は，*word₁* と *word₂*が同じウェブページに出現するウェブ検索ヒット件数を指す．

$$PMI\text{-}\mathrm{IR}(word_1, word_2) = log_2\left\{\frac{hits(word_1 \quad \& \quad word_2)}{hits(word_1)hits(word_2)}\right\} \qquad (3)$$

このように，PMI-IR は 2 つの表現の関連度を算出するものである．Turney の研究では，レビューサイトのコメントを”excellent”(ポジティブ)か”poor”(ネガティブ)の 2 極に自動分類する研究を行った．PMI-IR によって，コメントと”excellent”，または”poor”との関連度を算出し，どちらに強く関連しているかによって，判定を行っている．

この PMI-IR は，日本語のテキストにも効果があることが報告され[8]，また様々な分野でも適応されている[9]．本研究では，2 つの表現を掲示板の書込みと有害な極性を持つ単語との関連度の強さを算出するように拡張した．以下，拡張 PMI-IR 手法について述べる．拡張 PMI-IR 手法は，(１) 掲示板の書込みから phrase を抽出，(２) 極性単語の参照，(３) 拡張 PMI-IR による有害表現判別，という三つのステップで処理を実施する（図 6. 1）．以下，各ステップについて説明する．

6.1　　phrase の抽出

有害情報書込みにも無害な表現が含まれ，無害情報にも「鬼」や「怖い」などの有害の可能性を含む表現が含まれ，それぞれ有害無害表現が混在している．よって，判別要素には必要最小限の表現を抽出し，判別のノイズとなる要素を除去しなければならない．そこで，書込みから有害情報候補となる単語列（図 6. 1 では”phrase”と表記）を抽出する．

phrase には，有害情報の出現パターンに該当する単語列が判別のノイズが少なく，かつ有害情報の根本となる表現なので妥当だと考えられる．そこで，5.2 節で作成した標準化した書込み 2998 件(有害情報書込み 1508 件，無害情報書込み 1490 件)を対象に，有害情報の出現パターンを分析した．

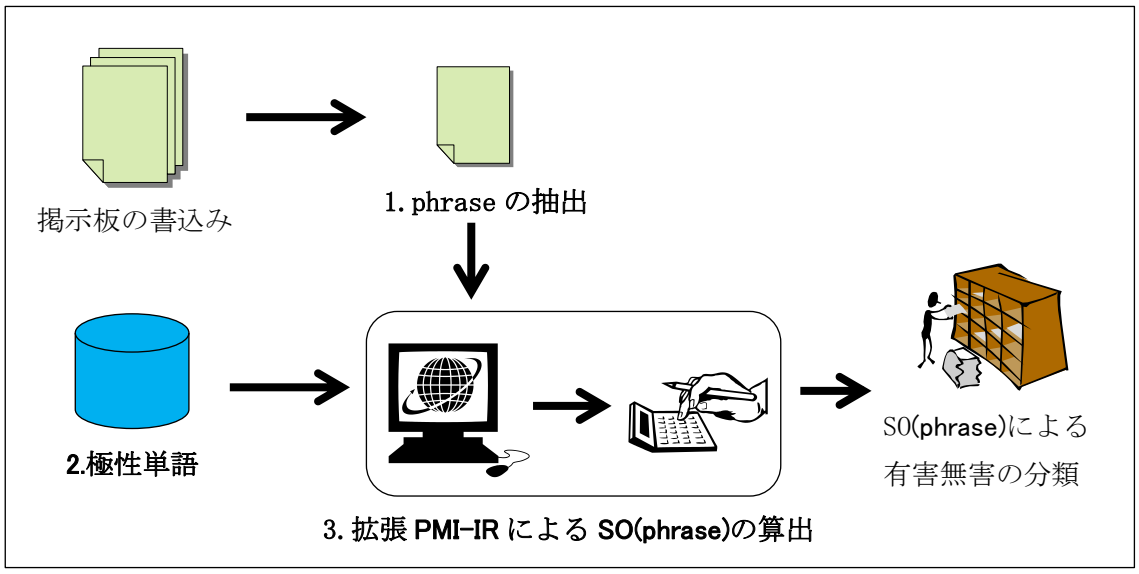

図 6. 1 提案手法の処理の流れ

6.1.1 有害情報の出現パターン

有害情報の出現パターンの分析を行った．まず，対象の書込みを形態素解析する．そして，有害無害間で重複している形態素を除去し，品詞別[2]に出現頻度順ランキングを作成する．そのランキングを調査した結果，名詞，動詞，形容詞で有害と無害の上位を占める単語に差異が見られた．それぞれの品詞別出現頻度順ランキングの一部を表 6. 1 に示す．表 6. 1 のように，有害の出現頻度順ランキングには有害単語が上位を占めており，無害の出現頻度順ランキングは様々なトピックの一般的な単語が占めていた．

表 6. 1 出現頻度順ランキングの一部

	名詞(有害)	名詞(無害)		動詞(有害)	動詞(無害)		形容詞(有害)	形容詞(無害)
1 位	パンコ	ラーメン		死ねる	行う		きもい	美味い
2 位	調子	会社		しねる	焼く		うざい	寒い
3 位	(人名)	味		殺す	生きる		キモイ	おいしい
4 位	(人名)	うどん		しぬ	頑張る		ウザイ	少ない
5 位	不細工	豆腐		やれる	切る		臭い	楽しい
6 位	ヤリマン	トンネル		殴る	売れる		気持ち悪い	面白い
7 位	先生	ケーキ		遊ぶ	集まる		キモい	旨い

また，有害の出現頻度順ランキングの内容を見ると，図 6. 2 のようになる．それぞれの品詞でみられる形態素に傾向が現れていた．各品詞の形態素出現傾向は，以下のようにな

[2] 連体詞，接頭詞，名詞，動詞，形容詞，副詞，接続詞，助詞，助動詞，感動詞

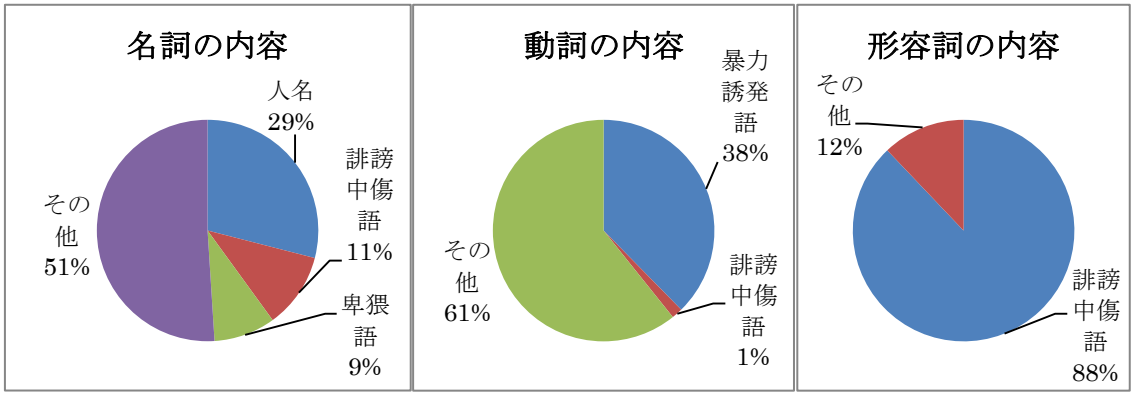

図 6.2 有害情報における品詞別の内容

っていた.

- 名詞には，人名や「バカ」などの誹謗中傷語や卑猥語が見られる.
- 動詞には，「死ね」などの暴力誘発語が見られる.
- 形容詞には，「キモイ」などの誹謗中傷語が見られる.

これは，先ほど述べた文部科学省の「サイト・スレッドの書込み類型化」に対応する．以上から，有害と無害間に分ける差異は単語の名詞，動詞，形容詞に現れ，それらの内容は文部科学省の類型化に対応することが分かった.

さらに，有害情報の係り受け関係を調べたところ，以下のように特定の要素が組み合わされるという条件によって有害化する傾向が見られた.

- 係り受け関係にある名詞と名詞の組
  (例)ゴリラ風の顔 → **「対象をある名詞で例える」**
  (例)ヤリマンの松葉 → **「対象の修飾」**
- 係り受け関係にある名詞と動詞の組　(例)松葉を殺す → **「対象に行動する」**
- 係り受け関係にある名詞と形容詞の組　(例)性格が悪い → **「対象を表現する」**

例えば，「性格が悪い」や「胸がでかい」などの有害表現は，「性格−悪い」「胸−でかい」という係り受けで構成されている．しかし，その構成要素である「性格」や「胸」，「悪い」，「でかい」のみが単独で出現したとしても有害性を持たず，これらの要素が係り受け関係を持って共起することによって，はじめて有害性を持つのである．よって，このような係り受け関係を持つ要素の組を有害性判定の素性として用いることは，有害情報の判別に大きく寄与すると考えられる．しかも，この係り受け関係にある形態素組は，「キモイ−ガイジ」などのようにそれ単体だけでも有害性を持つ表現も含んでいる.

6.1.2 phrase の抽出規則

phrase の抽出規則には，6.1.1 節の分析結果によって得られた係り受け関係を利用した．まず，入力として与えられた書込み文を係り受け解析する．係り受け解析の結果から，「名詞–名詞」，「名詞–動詞」，「名詞–形容詞」のいずれかの係り受け関係を持つ形態素の組を phrase として抽出，保持する．phrase の抽出例を表 6. 2 に示す．

表 6. 2 phrase の抽出例

抽出対象書込み　：可愛いけど性格が悪い女 抽出した phrase　:「可愛い－女」，「性格－悪い」，「悪い－女」

6.2 極性単語

ここでは，6.1 節の phrase との関連度を算出する，極性単語に選択した単語について説明する．極性単語とは，有害な極性単語なら有害情報にだけ頻出する，無害な極性単語なら無害情報にだけ頻出するような，偏った特性を持つ単語と定義する．PMI-IR を考案した Turney の研究では，ポジティブ(excellent)とネガティブ(poor)の 2 極性を持つそれぞれの単語との関連度を算出している．有害性の判定においても同様に，有害と無害の 2 極性になる．有害な極性単語は，疑わしいものは全てチェックという観点から有害単語を極性単語とすればよい．しかし，無害な極性単語は「面白い」という形態素を例に挙げると，「面白い髪形」と「面白いゲーム」のように，明確に無害だと判別可能な極性単語はほとんど無いと考えられる．

そこで，極性単語は有害な極性を持つ単語のみとした．極性単語には，有害情報に頻出する単語は有害性が高いと仮定し，表 6. 1 の有害情報の形態素出現頻度順ランキングにおける，卑猥語，暴力誘発語，誹謗中傷語に該当する上位 3 件ずつとした．極性単語を以下に示す．

- **卑猥語**　　:セックス，ヤリマン，フェラ
- **暴力誘発語**　:死ね，殺す，殴る
- **誹謗中傷語**　:きもい，うざい，不細工

6.3 拡張 PMI-IR による SO(phrase)の算出

以上の二つの要素から，意味の方向性：SO(phrase)を式(4)のように算出する．つまり，SO(phrase)は phrase と極性単語との関連度を示しており，値が大きければ大きいほど有害性が高く，値が小さければ小さいほど有害性は低いと考えられる．

$$SO(phrase) = \sum_{word \in 極性単語} PMI\text{-}IR(phrase, word) \quad (4)$$

また，一つの書込みには複数の phrase が存在する場合もあるので，複数の SO(phrase)が得られる．疑わしいものは全てチェック対象という観点から，SO(phrase)が最大のものをその書込みの SO(phrase)とした．

第 7 章　　評価と考察

提案した手法を用いて有害情報と無害情報の分類を行った．また，併せて比較実験も行った．以下に，比較実験について詳述する．

7.1　　比較実験の設定

比較実験として，SO(phrase)を算出する際に 4 種類の手法を加えた実験を用意した．さらに，市販されている有害情報フィルタリングソフトを想定し，有害単語マッチングによる有害情報の抽出実験も行った．以下に各実験について述べる．

7.1.1　4 種類の手法

1.前方参照手法　「する」や「の」などの,それ単体では意味が不十分な形態素の場合，2 個前まで形態素を取得する(例:する→抗議をする).

2.品詞別手法　図 6. 2 の分析結果を利用し，「名詞－名詞」の phrase には卑猥語と誹謗中傷語の極性単語を，「名詞－動詞」の phrase には暴力誘発語の極性単語，「名詞－形容詞」の phrase には誹謗中傷語の極性単語の組み合わせで SO(phrase)を算出する．

3.有害単語辞書手法　有害単語辞書は，5.1 節で取得した有害単語 239 語で，phrase に有害単語が含まれる場合に重み付けを行う(5).

4.単語感情極性辞書手法　単語感情極性辞書は，高村らが構築した辞書である[10]．単語には-1～+1 の値を付与されており,-1 に近いほどネガティブ,+1 に近いほどポジティブとしている.phrase に単語感情極性辞書の単語が含まれる場合に重み付けを行う(5).

$$SO(phrase) = log_2\left\{\frac{hits(phrase\ AND\ 死ね)}{hits(phrase)hits(死ね)} \times \alpha\right\} + \cdots \quad (5)$$

$\alpha = 1 + (有害辞書単語を含むなら + 1 \quad or$
$感情極性辞書で度合が - 0.99以下の単語を含むなら + 1)$

実験データセットとして 5.2 節で述べた修正後のデータ 2998 件(有害情報書込み 1508 件，無害情報書込み 1490 件)を用いた．全ての書込みを SO(phrase)順にランキングし，上位 n

件以上の書込みを有害情報と判定する，閾値 n を設定した．評価方法は，適合率(6)と再現率(7)で行う．また，上述した 4 種類の設定を加えていない実験を実験 1 とし，設定を加えた実験を実験 2 とする．

$$適合率 = \frac{閾値以上の有害情報書込み件数}{閾値以上の書込み件数} \quad (6)$$

$$再現率 = \frac{閾値以上の有害情報書込み件数}{全ての有害情報書込み件数} \quad (7)$$

7.1.2 有害単語マッチング手法による有害無害の分類

現在，市販されている有害情報のフィルタリングには「Yahoo!あんしんねっと」[3]や「ファミリーガード」[4]などがある．これらのフィルタリングソフトは，あらかじめ取得した有害なキーワードが，訪問サイトに含まれるかどうかによってフィルタリングしている．このようなキーワードフィルタリングソフトを想定し，有害単語マッチングによる有害情報と無害情報の分類を試みた．

有害単語には，5.1 節で取得した 239 単語を使用した．有害単語が書込み内に含まれていれば有害，含まれていなければ無害と判定する分類実験を行う．実験データセットは，7.1.1 節と同様である．評価方法は，適合率(8)と再現率(9)で行う．

$$適合率 = \frac{システムが有害と判定した書込み中の有害情報書込み数}{システムが有害と判定した書込み数} \quad (8)$$

$$再現率 = \frac{システムが有害と判定した書込み中の有害情報書込み数}{全ての有害情報書込み件数} \quad (9)$$

7.2 実験結果

まず，拡張 PMI-IR 手法による分類実験結果について述べる．phrase を抽出した結果，有害情報は 1508 件中 1034 件，無害情報は 1490 件中 1022 件抽出できた．phrase を抽出できなかった書込みは，表 7. 1 のように，一言だけの書込みや，意味不明の書込み，6.1 節の phrase 抽出規則に当てはまらない書込みだった．phrase を抽出できた 2056 件の書込みを対象に，実験 1，実験 2 を行ったところ適合率が図 7. 1，再現率が図 7. 2 がとなった．

そして，有害単語マッチングによる実験結果は，適合率 0.91，再現率 0.42 となった．

[3] http://anshin.yahoo.co.jp/
[4] http://technos-jp.com/family/

表 7. 1 phrase を抽出できなかった書込み

・きもい・・・
・njdcn
・キモイ，本当に死ねばいいとおもう

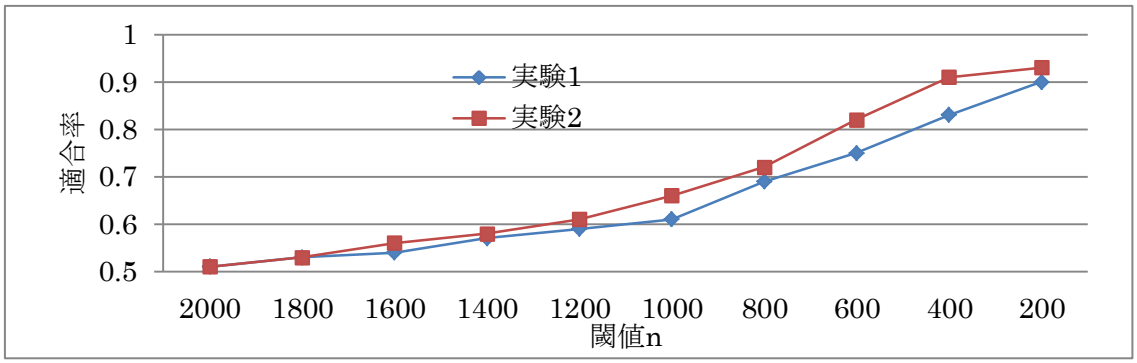

図 7. 1 拡張 PMI-IR による実験の適合率

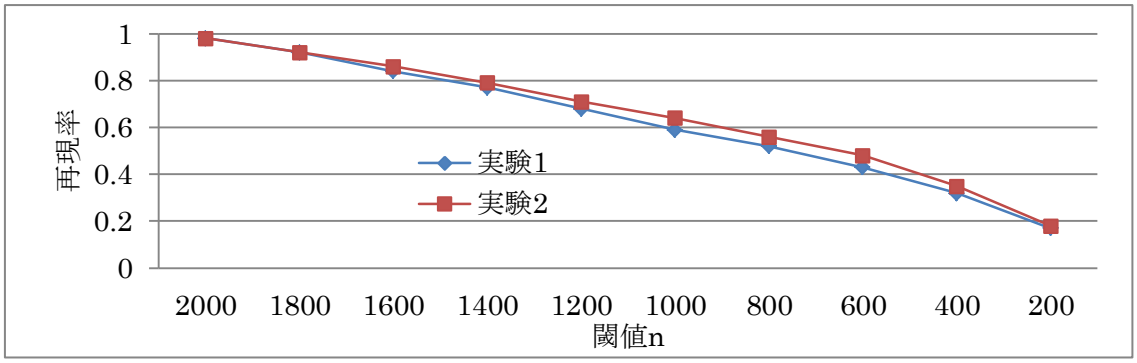

図 7. 2 拡張 PMI-IR による実験の再現率

7.3 考察

まず，拡張 PMI-IR 手法による実験結果について考察する．phrase を SO(phrase)順ランキングした結果(表 7. 2)を基に考察した．上位の phrase を調べると，有害情報では「きもいー顔」や「ヤリマンー女」などの極性単語が含まれる phrase が多く占めている．さらに，「しゃくれー男」や「黙れーガイジ」などの極性単語に選択しなかった有害単語を含む phrase も上位に位置していた．つまり，拡張 PMI-IR 手法では極性単語以外の未知語，新語に該当する有害単語も分類でき，「うざい」が「うっざい」のように少しだけ変えた，日々増え続ける表現にも対応できるということである．この結果から，有害情報を含むウェブ

ページは，あまり使用されない有害単語のみで構成されることは無く，頻出する「死ね」などの有害単語が少なからず含まれるものだと推測される．また，「眉毛ー濃い」や「乳ー

表 7. 2 phrase の SO(phrase)順ランキング

	有害			無害
1 位	おまんこーくさい		1 位	殺人ー自殺
2 位	”人名”ー男たらし		2 位	給料ー下がる
3 位	”人名”ーパンコ		3 位	地震ーない
4 位	デブーブス		4 位	のどー痛い
5 位	きもいー顔		5 位	のどー痛い
6 位	顔ーきもい		6 位	トンネルー怖い
7 位	うざいー先輩		7 位	奴ーかわいく
⁀	⁀	⁀	⁀	⁀
1028 位	”人名”ー ”人名”		1016 位	感ーいい
1029 位	6ー玲		1017 位	西ー空
1030 位	年ー子		1018 位	食品ー0
1031 位	史ー中		1019 位	2ー・
1032 位	年ー5		1020 位	1ー・
1033 位	年ー章		1021 位	高ー年
1034 位	殿中ー人		1022 位	山之一色ー仙

でかい」などの単体では無害な phrase も上位に位置していた．逆に，下位の phrase は「松葉ー達明」や「史ー中」などの，悪口を含まない人名や，メールアドレスなどの個人情報の流布や，意味の取れない phrase などが占めていた．無害情報では，上位の phrase は「あそこー美味い」や「のどー痛い」などが占めていた．これは，「あそこ」は指示代名詞であり，卑猥語と共起が高くなったり，「痛い」は暴力誘発語と共起が高くなってしまったからである．また，表 7. 2 の無害情報 1 位に「殺人ー自殺」という，有害情報のように思われる phrase がある，しかし，これは「>>53, ありがとう　なんか事件でもあったんですかね？殺人とか自殺とか」という無害情報の書込みから抽出できた phrase である．今回は，筆者が無害と判定したが，人によっては有害と判断するかもしれない曖昧な書込みである．よって，拡張 PMI-IR 手法では，このような曖昧な書込みは上位に位置付けられ，疑わしいものは全てチェックできることが分かった．下位の phrase は，「ジンギスカンー食べ」や「食品ー0」などの無害なものや，意味の取れないものであった．

比較実験に設定した実験 2 では，どの閾値においても実験 1 より分類性能が良かった．そこで，実験 2 で追加した設定を一つ一つ実験しなおし，それぞれの設定がどのような効

果をもたらしたのか分析を行った．まず，前方参照手法において，有害情報では「先生－し」という組が「先生－ワイセツ行為し」のようになり，有害情報と判定できるようになっていた．しかし，無害情報では，「今月－さ」というのが「今月－首宣告さ」のように極性単語と共起の強い形態素が現れ，SO(phrase)を引き上げてしまっていた．この結果から，極性単語に選んだ単語が有害無害どちらにも関連の強い語を選んでいる可能性がある．品詞別手法は，あまり効果がなかった．このことは，有害情報の類型は有害表現を規定する有効な軸とはなり得ないことを示唆している．単語感情極性辞書による重み付けは，品詞別手法と同じく効果がなかった．これは，ポジティブ／ネガティブの極性が必ずしも有害性の判定には寄与しないことを示している．例えば，「事故」という言葉はネガティブな言葉だが，有害な方向にのみ強く関連する単語ではない．一方，有害／無害の極性軸上にある，有害単語辞書による重み付けは大きな効果があった．つまり，一般的なポジティブとネガティブの極性は，有害と無害の極性と性質が大きく異なると考えられる．

次に，有害単語マッチング手法による実験結果について考察する．適合率は，0.91 であり，抽出失敗には「行政に優遇されてばかりじゃないですか」のように「ばか」を過抽出してしまっていた書込みだった．再現率は，0.42 であり，抽出できなかった有害情報には，拡張 PMI-IR 手法と同様に住所や電話番号などの個人情報，人名などが取れておらず，さらに「性格－悪い」のような組み合わせによって有害化する表現が取れていなかった．

拡張 PMI-IR 手法と有害単語マッチング手法の考察をまとめると以下のようになる．

1. 拡張 PMI-IR 手法では，「性格悪い」などの組み合わせによって有害化する表現を含む有害情報が取れる．さらに，極性単語以外の有害単語を含む有害情報を取れることから，表記ゆれが起こった有害単語や新しい有害単語にも対応できる．
2. 有害単語マッチング手法でしか取れない有害情報は，「死ね」や「キモイ，本当に死ねばいいとおもう」などの一言や phrase 抽出規則に当てはまらない有害情報が該当する．しかし，一言だけの書込みは書込み自体を phrase とするなど，phrase の抽出規則を拡張すれば対応できると考えられる．
3. どちらの手法でも取れない有害情報は，「asdf@docomo.ne.jp マツバのメールアドレス」や「三重大学大学院松葉達明」のような個人情報の流布，人名などの有害情報が該当する．しかし，このような個人情報の取得は先行研究[11]が行われており，この研究成果を応用すれば対応できると考えられる．

第 8 章　　結論

「ネット上のいじめ」における有害情報の分析を行い，PMI-IR を応用して有害無害の分類を試みた．まず，形態素の出現頻度順ランキングを作成し分析を行った．結果，有害情報の名詞には人名や誹謗中傷，卑猥な語が見られ，動詞には「死ね」などの暴力を誘発する語，形容詞には「キモイ」などの誹謗中傷語が頻出することが分かった．また，係り受け解析による分析も行ったところ，「名詞と名詞」，「名詞と動詞」，「名詞と形容詞」の組において有害情報の出現パターンがあることが分かった．そして，拡張 PMI-IR による有害無害の分類を試みた．分類実験の結果，誹謗中傷による有害情報は分別可能だが，人名や個人情報は分類が難しいことが分かった．また，拡張 PMI-IR 手法では「性格悪い」などの組み合わせによって有害化する有害情報や，極性単語以外の未知語，新語に該当する有害単語を含む有害情報も分類できることが分かった．比較実験の結果からは，一般的なポジティブとネガティブの極性は，有害と無害の極性と性質が違うことが分かった．さらに，市販されている有害情報フィルタリングソフトを想定し，有害単語マッチング手法による比較実験を行った．その結果，有害単語マッチング手法でしか取れない有害情報はあるが，拡張 PMI-IR 手法の phrase の抽出規則を拡張すれば対応できるものだと分かった．また，両方の手法で取れない個人情報の流布などは，個人情報の抽出に関する先行研究がなされており，その研究を応用すれば取得できると考えられる．

謝辞

本論文に関する研究を進め，論文を完成させるにあたり，本当に多くご指導とご支援を賜りました，北見工大の桝井文人准教授，河合敦夫准教授，井須尚紀教授に深く感謝致します．本研究を進めるにあたり，学校非公式サイト情報を提供頂いた財団法人反差別人権研究所みえの松村元樹研究員に深く感謝致します．お忙しいところ，副査を御引受け下さった佐々木敬泰助教に深く感謝致します．いろいろと便宜を図っていただきました吉永みゆきさんに深く感謝致します．最後に，楽しい研究生活を共にした，研究室の皆様に深く感謝致します．

参考文献

[1] 文部科学省.「ネット上のいじめ」に関する対応マニュアル事例集(学校・教員向け).文部科学省,2008.

[2] 渡辺凡,砂山渡:電子掲示板におけるユーザの性質の評価.電子情報通信学会技術研究報告.No.652 in 2006-KBSE,pp.25-30,2006.

[3] 石坂 達也,山本 和英: 2 ちゃんねるを対象とした悪口表現の抽出.言語処理学会第 16 回年次大会, pp.178-181 (2010.3)

[4] 池田和史, 柳原正ら: 格要素の抽象化に基づく違法・有害文書検出手法の提案と評価. 情報処理学会第 72 回全国大会,pp.71-72(2010.3)

[5] 形態素解析システム茶筌：http://chasen-legacy.sourceforge.jp

[6] 日本語係り受け解析器南瓜：http://chasen.org/～taku/software/cabocha/

[7] Peter D. Turney：Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp.417-424(2002,7)

[8] Guangwei Wang,Kenji Araki: Modifying SO-PMI for Japanese Weblog Opinion Mining by using a balancing factor and detecting neutral expressions. In Proceedings of the North American Chapter of the Association for Computational Linguistics, pp. 189-192(2007).

[9] 深澤佑介, 太田順 : PMI-IR の拡張による Web からのユーザ行動間の関連性抽出. 人工知能学会全国大会（第 24 回）,3B4-3(2010.6).

[10] 高村大也, 乾孝司, 奥村学：スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp. 627-637, 2006.

[11] 浅野久子, 加藤恒昭ら：Signature の局所的パターンマッチによる電子メールからの送信元住所録情報の抽出とそれを用いた住所録管理システム.情報処理学会論文誌 39(7), pp.2196-2206, 1998-07-15